PHIL 7001: Fundamentals of AI, Data, and Algorithms

Week 5A
Inference

Boris Babic,
HKU 100 Associate Professor of Data Science, Law and Philosophy

Learning Goals

- Understand the concepts of statistical inference, estimation, population, and sample.
- Explore the fundamentals of hypothesis testing.
- Understand null and alternative hypotheses.
- Grasp the significance level and p-value in hypothesis testing.
- Apply hypothesis testing to a practical example.

Review From Last Class

Last Class

Review of last class

- Probability distributions
- Binomial distribution
- Normal distribution
- Using R to calculate binomial and normal probabilities

Practice Question 1

- In a population, 30% of people have a certain genetic trait. If you randomly sample 200 people, what's the probability that fewer than 60 of them have the trait?

Solution to Practice Question 1

- In a population, 30% of people have a certain genetic trait. If you randomly sample 200 people, what's the probability that fewer than 60 of them have the trait?
- Using pbinom():

$$prob \leftarrow P(X \leq 59) = \text{pbinom}(59, size = 200, prob = 0.30)$$

- Using dbinom() for individual probabilities and summing:

$$prob \leftarrow sum(dbinom(0:59, size = 200, prob = 0.3))$$

Practice Question 2

- A factory produces light bulbs, and 95% of them are functional. If you randomly select 50 light bulbs, what's the probability that at most 3 of them are non-functional?

Solution to Practice Question 2

- A factory produces light bulbs, and 95% of them are functional. If you randomly select 50 light bulbs, what's the probability that at most 3 of them are non-functional?
- Using pbinom():

$$prob \leftarrow \text{pbinom}(3, size = 50, prob = 0.95)$$

- Using dbinom() for individual probabilities and summing:

$$prob \leftarrow sum(dbinom(0:3, size = 50, prob = 0.95))$$

Practice Question 3

- Given a random variable following a normal distribution with mean 200 and standard deviation 25, use R to find the probability that it takes on a value between 180 and 220.

Solution to Practice Question 3

- Given a random variable following a normal distribution with mean 200 and standard deviation 25, use R to find the probability that it takes on a value between 180 and 220.
- To solve this question, we can use subtraction:

$$prob \leftarrow pnorm(220, mean = 200, sd = 25) -$$

$$pnorm(180, mean = 200, sd = 25)$$

Statistical Inference

Population and Sample

- **Population** is the set representing all measurements of interest to the investigator.
- A population can be any entire collection of people, animals, plants, or things from which we may collect data. It is the entire group we are interested in, which we wish to describe or draw conclusions about.
- In order to make any generalizations about a population, a **sample**, that is meant to be representative of the population, is often studied.
- A sample is a subset of measurements selected from the population of interest.

Statistical Inference

- Statistical Inference makes use of information from a **sample** to draw conclusions about the **population** from which the sample was taken.
  - We use probability tools to make educated guesses about population parameters.
- It mainly consists of two parts:
  - Estimation
  - Hypothesis Testing
- Terminology:
  - **Parameter**: A quantity describing the population. This is something we don't know, but wish to know. For example: the true prevalence of Covid. Let's call it p.
  - **Statistic**: A measure calculated from the sample data. This is something we can observe. For example, the prevalence of Covid in a sample of 100 randomly selected people. Let's say it is 24. Our sample mean in this case is 0.24.
  - Inference: A way of making a guess about the unknown parameter, using the observed statistic.
  - A naive guess: if the sample is 24/100 (i.e., 0.24) then the true parameter is also 0.24.

Statistical Inference

	Parameter	Statistic
Size	N	n
Mean	μ	x̄
Standard deviation	σ	s
Proportion	P	p
Correlation coefficient	ρ	r

Estimation

- Estimation is the process whereby we select a random sample from the population and use a sample statistic to estimate the population parameter
- Two ways of estimation:
  - Point estimate
  - Interval estimate
    - Confidence Interval

Point Estimate

- **Point estimate** - A sample statistic used to estimate the population parameter
- Example: Using a sample mean ($\bar{x}$) as an estimate of the population mean (μ).
  - Scenario: We want to estimate the average income of people in HK.
  - We can't possibly survey the entire population.
  - Sample Mean ($\bar{x}$): The average income of a sample of 1000 randomly selected HK residents.
  - A naive estimation procedure: $\mu = \bar{x}$.
  - This is what we also did above in the Covid example.

Hypothesis Tests

Hypothesis Test

- Hypothesis tests are statistical methods for making decisions or inferences about population parameters.
- Purpose: Determine whether an assumption or belief about a population is supported by data.
- Hypothesis tests are evaluated using probability, but we have to be careful to calculate the right quantity. It is easy to get confused.
  - Null Hypothesis
  - Alternative Hypothesis

Types of Hypothesis

- Null Hypothesis (H_0)
  - The null hypothesis, denoted as H_0, represents a statement of no effect, no difference, or no change. It is often a statement of the status quo or a default assumption.
  - The null hypothesis is used as a starting point for the hypothesis test. It serves as a reference point for comparing the observed data or sample statistics.
- Alternative Hypothesis (H_1 or H_a)
  - The alternative hypothesis, denoted as H_1 or H_a, represents a statement that contradicts or opposes the null hypothesis. It is what the researcher is trying to find evidence for.

P-value and the significance level

- The **p-value** is defined as the probability of our result, or a more extreme result, under the null hypothesis.
- Historically, it was a rough and ready way to get a sense of the weight of evidence. It has now become a rule for evaluating whether one has a statistically significant result.
- A small p-value suggests strong evidence against the null hypothesis.
- The **significance level**, denoted as α, is the probability of making a Type I error (incorrectly rejecting a true null hypothesis).
- Strategy: Reject the null hypothesis if the p-value is less than the level of significance α.
- Conventionally, the significance level is always set to 0.05 (even though there is no reason for this!).
- So: to perform a hypothesis test: we have to state the null hypothesis, and then calculate the p-value under the assumption that the null hypothesis is true. If this p-value is smaller than 0.05, then the probability that the null hypothesis is true is very law. Therefore, we can reject the null hypothesis.
- Does this sound confusing to you? If it does, that is because it IS confusing. Hypothesis testing is a very strange way of performing statistical inference.

Example

A Curious Coin

You have come across a curious coin. It seems (you suspect) bent in a way that biases it toward landing on heads. You will give this coin to your trusty RA, and ask them to perform an experiment (i.e., toss it repeatedly) in order to help you decide whether the coin is biased.

Example

- Null hypothesis: the coin is fair. We can write this as $\theta = 0.5$.
- I have used the symbol theta here to denote a proportion, instead of the letter p, so that we do not create confusion between the p-value, and the proportion parameter.
- Often, we will use θ to denote proportions. So for the binomial distribution, we will often write $\binom{n}{x}\theta^x(1-\theta)^{n-x}$.
- Alternative hypothesis: $\theta \neq 0.5$.

Hypothesis Testing

- Formulate Hypotheses (H_0 and H_1). We have done this.
- Collect Data (e.g., toss the coin multiple times). We will do this soon.
- Analyze Data Using Statistical Tests (e.g., calculate p-value). We will do this soon too.
- Make a Decision Based on the Test Result (e.g., reject or fail to reject H_0). This is our conclusion.

Some Data

A Curious Coin

Your RA reports having tossed the coin 12 times, with the following results:

$$H, T, H, H, H, H, H, T, H, H, H, T$$

$$9H, 3T$$

Example

- Recall:

$$H_0 : \theta = 0.5, H_1 : \theta \neq 0.5.$$

- From our experiment,

$$H, T, H, H, H, H, H, T, H, H, H, T$$
$$9H, 3T$$

- How should we compute the probability of observing our result, or a more extreme result, under $H_0 : \theta = 0.5$? What is more extreme than 9H and 3T is we assume that $\theta = 0.5$?
- First: 9H, 3T + 10H, 2T + 11H, 1T + 12H, 0T
- Second: 3H, 9T + 2H, 10T + 1H, 11T + 0H, 12T
- We need to calculate: First + Second.
- This is easy for us to calculate in R.
- First: pbinom(8, 12, 0.5, lower.tail=FALSE) = 0.07.
- Second: pbinom(3, 12, 0.5, lower.tail=TRUE) = 0.07.
- Therefore: p-value = 0.07 + 0.07 = 1.4. The result is not statistically significant. Therefore, we cannot reject the hypothesis that the coin is fair.

Today

- Understand the concepts of statistical inference, population, and sample.
- Explore the fundamentals of estimation and hypothesis testing.
- Understand null and alternative hypotheses.
- Grasp the significance level and p-value in hypothesis testing.
- Apply hypothesis testing to a practical example.

Next Class

- Understand Bayesian inference
- Apple Bayesian inference to a practical example.